ELSEVIER

Geoderma 93 (1999) 177–194

GEODERMA

Quantification and simulation of errors in categorical data for uncertainty analysis of soil acidification modelling

P.A. Finke [a,*], D. Wladis [b], J. Kros [a], E.J. Pebesma [c], G.J. Reinds [a]

[a] *DLO Winand Staring Centre for Integrated Land, Soil and Water Research (SC-DLO), P.O.Box 125, 6700 AC Wageningen, Netherlands*

[b] *Department of Geology, Chalmers University of Technology, S-412 96 Gothenburg, Sweden*

[c] *Faculty of Environmental Sciences, University of Utrecht, P.O.Box 80115, 3508 TC, Utrecht, Netherlands*

Received 9 June 1998; received in revised form 1 June 1999; accepted 5 July 1999

Abstract

Simulation studies that use maps to generate georeferenced model input may be prone to errors in the definition and delineation of the map units. Our study aims at the estimation of errors in categorical data, i.e., a generalized soil and vegetation class map of the EU vs. a highly detailed soil/vegetation map of the Netherlands. From this, an error model evolves containing (i) an index of map purity and misclassified area fractions and (ii) indicator variograms describing the spatial autocorrelation structure of the degree of error at individual locations. Furthermore, we describe a method to evaluate the effect of these errors on the uncertainty of the outcome of the soil acidification model Simulation Model for Acidification's Regional Trends, version 2 (SMART2). This method involves the application of joint sequential simulation to produce equiprobable realisations of soil/vegetation maps. Results show that the errors in the EU-soil/vegetation map are considerable, because 69% of the area is misclassified when compared to highly detailed maps from the Netherlands. Simulated maps reproduced the error model for the dominant soil/vegetation map units well. Results of the uncertainty analyses show that errors in categorical data do have a pronounced influence on the uncertainty of SMART2 results. This influence was between 20% of the total variance for Al^{3+} concentrations and exceedance probabilities, and 40%–50% of the total variance for NO_3^- concentrations and exceedance probabilities.

Keywords: geostatistics; acidification; variance analysis; soil maps; vegetation

* Corresponding author. Fax: +31-317-424812.
E-mail address: p.a.finke@sc.dlo.nl (P.A. Finke)

0016-7061/99/$ - see front matter
PII: S0016-7061(99)00056-7

1. Introduction

In recent decades, increased knowledge of the behaviour of the soil–water–atmosphere system has led to the development of many computerized models. The basic uses of these models are to provide understanding of the current system status as well as to forecast future trends. Some of these models are the implementation of quantitative-mechanistic knowledge of relevant processes, while others contain qualitative-empirical expressions summarizing the net effect of a group of processes (Hoosbeek and Bryant, 1992). Usually, the type of approach chosen depends upon the spatial and temporal scale of application, the associated data availability and the knowledge of the system. De Vries et al. (1998) treat this issue in a family of soil acidification models involving i.a. the simulation Model for Acidification's Regional Trends (SMART).

Simulation Model for Acidification's Regional Trends, version 2 (SMART2) predicts the response of the soil solution concentrations below the root zone of seminatural vegetation due to changes in atmospheric deposition. The most important soil solution constituents that react to atmospheric deposition are, from an environmental viewpoint, nitrate (NO_3^-) and aluminium (Al^{3+}). SMART2 is a vertical one-layer model which is typically run on a point support, because (i) most input data are available for this support; (ii) upscaling of input data to the regional target scale is unwise since the model is not strictly linear to all its inputs, (iii) using this approach the distribution of point concentrations within blocks is obtained as well, and (iv) comparisons to (point) measurements can easily be made. SMART2 results always relate to seminatural vegetation and are usually transferred to the regional scale by upscaling 25-point support model outputs within each 5 km $\times$ 5 km block to 1 aggregated block median value. SMART2 needs input of 18 continuous soil and vegetation-related parameters, which are a function of soil and vegetation type (Kros et al., 1999).

The application of any model at any scale yields predictions with an associated error. Quantification of this error is important when the model is used to forecast future trends for different scenarios, because it determines the relevance and significance of the differences in the outcomes of two scenarios. In a deterministic system, the uncertainty in basic data and errors in the structure of the model itself are the two causes of prediction uncertainty (Jansen, 1998), and together they determine the prediction error. The basic input data for the SMART2-model can be divided into categorical data (e.g., soil type, vegetation type) and continuous data (e.g., CEC, transpiration). Categorical data are often not direct input parameters to the model, but are used for stratification or estimation of the continuous data, e.g., soil hydraulic parameters may be estimated with different continuous pedotransfer functions in case of clayey and sandy soils (Wösten et al., 1995), or the average CEC value for a clay soil differs from that of a loess soil. This implies that errors in continuous parameters are subject to the categorical data too. A Monte Carlo analysis of uncer-

tainty due to basic data should therefore follow a nested approach. Typical steps are (i) estimation of errors in categorical data; (ii) simulation of realisations of categorical data; (iii) estimation and simulation of errors in continuous data for each realisation of the categorical data; (iv) analyses of variance, whereby the uncertainty contributions of both types of data are separated.

This paper focuses on the uncertainty of categorical input data to the soil acidification model SMART2. We limit ourselves to (i) estimation and expression of the uncertainty of the model input as long as it can be attributed to uncertainty in the categorical data, i.e., the underlying soil and vegetation maps, and (ii) simulation of the realisations of the categorical data which serves as part of the input for the uncertainty analysis. The method of the uncertainty analysis itself, as well as the results obtained, has been reported elsewhere (Kros et al., 1999) and will only be summarized for the sake of completeness.

2. Material and methods

2.1. Description of data sets

The uncertainty analysis concerned the EU-wide application of the SMART2 model, using the available data for this area. For the estimation of errors in the categorical data, both the available EU-wide data as well as highly detailed data from The Netherlands were used, which served as ''ground truth''. For the simulation of the realisations of categorical data (see below), EU data were used.

The original sources of the categorical data used on the EU-scale were (i) the digital 1:1,000,000 soil map of the EU (EC, 1985) with minimal polygon sizes of approximately 2500 ha, and (ii) the CORINE-landcover database (EC, 1993), with pixel sizes of 25 ha. The original data sources covering The Netherlands were (i) the digital 1:50,000 Soil Map of The Netherlands (Steur and Heijink, 1991), with minimal polygon sizes of approximately 6 ha, and (ii) the land cover database of the Netherlands, with a square pixel area of 0.0625 ha (Thunnissen et al., 1992).

These data were generalized into a limited number of classes and gridded to 1 km $\times$ 1 km cells (EU-data) and 25 m $\times$ 25 m cells (NL-data). The dominant soil type and vegetation type were set for each grid cell after classifying soil types and land cover classes into a generalized legend of seven soil classes and four vegetation types. The soil classes were: poor sand (PS), rich sand (RS), calcareous sand (SC), noncalcareous clay (CN), calcareous clay (CC), loess soils (LN) and peat soils (PN). The vegetation types relating to natural vegetations were: coniferous forest (CON), deciduous forest (DEC), heather (HEA) and

nonfertilized grassland (GRP). The SMART2 model utilised the combination of soil type and land use as categorical input data. Because it can be assumed that soil type and vegetation type are interdependent, the different categories were combined to unique categorical variables (Kros et al., 1999). By using these combinations for the simulations rather than each data type separately, the interdependencies between soil and land use categories were preserved. In total, 28 combinations of soil class and vegetation types were possible, both for the EU-data and the NL-data. From this point on, we will refer to the 1 km × 1 km EU soil–vegetation data as the *SV-EU map* and the 25 m × 25 m NL soil–vegetation data as the *SV-NL map*.

The purpose of the simulation was to produce realisations of the soil/land use map for the Netherlands suitable for modelling of nitrate and aluminium concentrations. Since the uncertainty of input data was to be taken into account as well, and the analysis was to be valid for the whole EU, data had to fulfil the following requirements (Kleeschulte, 1997): (i) the data must be available for the whole of Europe; (ii) the data must be harmonised according to a common nomenclature in order to avoid regional or national inconsistencies; (iii) the data should be available in a seamless database; (iv) the data should be available from one single source to avoid regional or national inconsistencies. This implies that the processed data should comply with the Eurostat-GISCO data format. Therefore, data were processed in 5 km × 5 km blocks on a regular grid.

2.2. Quantification of errors in categorical data (map impurity)

The categorical EU-data were based upon less detailed data sources than the NL-data, and therefore were expected to contain considerable more error than the NL-data. The working hypothesis of this study was that errors in the highly detailed NL-data were small compared to the EU-data, and therefore NL-data could be used as ground truth to estimate the errors in the EU-data. Also, the EU-maps were constructed independently (by different surveyors) of the NL-maps and are based on older surveys. The error in the detailed NL-maps is known to some extent. The fraction of the area occupied by a land cover type that actually corresponds to the classification (the map purity) is nearly 90% for the natural vegetations (Noordman et al., 1997). The target map purity of the soil map 1:50,000 is 70% (Steur and Heijink, 1991). This percentage is known to be less when the water table class (part of the definition of map units) is considered, but appears realistic when only parent material, calcareousness, soil type and texture class are considered, as in the current study. No data, however, were available to support this statement.

For convenience, we will refer to soil–vegetation combinations as *EU-categories* when they occur on the SV-EU map. The soil–vegetation combinations that appear on the SV-NL map within a SV-EU category will be referred to as

NL-classes. To quantify the degree of error within a EU-category, we introduced an indicator variable $I_{t,s}$ for NL-class t within EU-category s:

$$I_{t,s}(x) = \begin{cases} 1 & \text{if } \mathrm{EU}(x) = s \text{ and } \mathrm{NL}(x) = t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathrm{EU}(x)$ is the EU-category at location x and $\mathrm{NL}(x)$ is the ''true'' soil/vegetation class estimated by the NL-class at location x.

The agreement between EU-category s and NL-class t is then estimated by:

$$\hat{p}_{t,s} = \frac{\sum_{x=1}^{n_s} I_{t,s}(x)}{n_s} \quad (2)$$

where n_s is the number of sample locations within EU-category s and $\hat{}$ denotes an estimation. Within a GIS, the agreement p (or the disagreement, $1-p$) could also be estimated directly by: (i) overlaying SV-EU with SV-NL, and calculating the area in each EU-category attributed to each NL-class; (ii) expressing these areas as area fractions by division through the total area in the category attributed to any of the distinguished soil–vegetation units. Water, built-up areas and agricultural landuse were not considered in the analysis. The fractions thus obtained equalled the estimated values of $p_{s,t}$ of occurence of the above indicator variables, and could also be termed NL-class impurity contributions to one EU-category.

As a first indication of the uncertainty due to map inaccuracies, the distribution of $I_{t,s}$ needed to be known. Probability density functions of this type of indicator variable follow the binomial distribution, and the variance of $I_{t,s}(x)$ was therefore calculated as:

$$s^2(I_{t,s}(x)) = \hat{p}^*_{t,s}(1-\hat{p}_{t,s}) \quad (3)$$

2.3. Quantification of the spatial correlation structure of errors

The uncertainty analysis concerns expressing the uncertainty of model outputs at supports (blocks) exceeding those of SMART2. In this case, it was necessary to take the spatial correlation of model input parameters into account, to realistically assess the uncertainty due to within-block variability as well as that due to between-block variability (Kros et al., 1999). Since the current uncertainty analysis was based upon simulation of realisations, the spatial autocorrelation structure of the indicator variables needed to be estimated. The following procedure was followed.

(i) The map of an EU-category s was put into a separate coverage;

(ii) Within each cell of 1 km $\times$ 1 km in this coverage, 20 locations x were drawn at random;

(iii) In the coverage of the SV-NL map (25 m × 25 m cells), the actual soil type and vegetation (t) was looked up at each x;

(iv) $I_{t,s}(x)$ was calculated using Eq. 1;

(v) Assuming second-order stationarity, sample indicator variograms were determined for distances between 0 and 30 km, using x and $I_{t,s}(x)$ with the GSTAT-package (Pebesma and Wesseling, 1997);

(vi) Variogram models were fitted to the sample variogram, using a weighed least-squares criterion. After having analyzed the sample variograms, it was decided to fit an exponential variogram in all cases:

$$\gamma(h) = C_0\delta(h) + C_1(1 - e^{-h/A}) \quad (4)$$

where $\delta(h)$ is 0 when the distance h is 0, and 1 in other cases, C_0 is the nugget variance, $C_1 + C_0$ is the sill variance and A is a parameter related to the range.

Ideally, the sum of $C_1 + C_0$ should equal the binominal variance as estimated by the GIS-overlay. This does not occur in practice due to the following reasons: (i) the sample variograms are based upon a sample of 20 points/km^2, while the binominal variances are based on the whole population of 1600, 25 m × 25 m NL grid cells per km^2 EU, (ii) C_1 and C_0 result from a fitting procedure, and (iii) variograms are non-decreasing functions, and therefore the limiting variance of the variogram model must be larger than the estimated variance which is calculated, irrespective of the distance between observations. For these reasons:

(vii) C_1 and C_0 are scaled with a factor F, so that $F^*(C_1 + C_0) = s^2$ where s^2 is estimated by Eq. 3.

(viii) Steps (i) to (vii) are repeated for all other EU-categories.

The variogram model thus obtained, its parameters C_0, C_1 and A as well as the areal fractions p are together referred to as the *error model*.

2.4. Simulation of soil / vegetation maps

To match the objectives of the project, the algorithm selected for joint simulations should allow a truthful reproduction of the error model, i.e., the statistical properties. Therefore, the method employed to produce the realisations should fulfil the following prerequisites: (i) the error model is fully taken into account, and (ii) the spatial correlation of the error is incorporated.

By taking not only the misclassification for each class into account, but also how the misclassification is subdivided into all the remaining classes, the first prerequisite is met. The second prerequisite stems from the assumption that misclassifications appear in clusters rather than at isolated points (Kros et al., 1999) and is satisfied by incorporating their spatial correlation structure into the error model. The realisations of categorical data were produced using joint sequential simulation (Goméz-Hernández and Journel, 1992). For the simula-

tions, the GSTAT program (Pebesma and Wesseling, 1997) was used. The simulation procedure employed in this study disregards both spatial correlation between different classes within a category, and spatial correlation across EU-category boundaries. However, the simulation procedure employed herein lends itself perfectly well to handling any type of cross-covariance matrix (Goméz-Hernández and Journel, 1992), but this was considered to be beyond the scope of this study.

For the simulation of soil/land classes, a sequential multiple indicator simulation for categorical variables (Goovaerts, 1997, p. 423) is used. Conditioning on the true (NL) soil/land classes is obtained by using one unconditional (prior) probability vector and a set of variograms for each NL class, in order to reproduce spatial correlation only within NL-classes. During the simulation, only previously simulated data within the same NL-class is used as conditioning data. Multiple realisations were obtained by repeating this procedure independently. For further details on sequential indicator simulation of categorical variables, we refer to Goméz-Hernández and Journel (1992) and Deutsch and Journel (1992).

To produce a map of one realisation, the following steps were performed.

(i) For each of the 28 categories, or strata, the abundance of all classes within each specific stratum were simulated, yielding an indicator map for each class.

(ii) The indicator maps were converted to nominal classes by assigning a number to each typical combination of soil and vegetation class (i.e., coded 0–27).

(iii) The categories were combined to one single soil/vegetation map.

In the evaluation of the simulation results, the method for reproducing the error model (both the areal fraction of the classes and spatial correlation, i.e., the variograms) in the realisations was studied.

Simulations are conditioned if all realisations honour the hard data values at their locations. In this project, no hard data values in terms of field data were used. The use of such data was beyond the scope of this project, since for this, data should exist and be available on the European level. The methodology is thus also applicable in parts of Europe where no observation data are available. Therefore, the simulations were performed as unconditional in the sense that no conditional data in the form of point data were used.

Simulation was stratified, which implies that the simulation region is subdivided into subregions or strata. Each stratum, in this case an EU S/V map category, has its unconditional (prior) probability vector. By means of stratification, the known information was preserved. In the present study it implies that the boundary for each category is honoured. In this context, stratification can be seen as a way of conditioning known information, i.e., the areal extent of the categories.

For the evaluation of the simulated results, we quantified how the error model was reproduced in the realisations. The areal fraction of different classes within

each stratum was studied and compared to the error model. The spatial correlation of the simulated categories was studied by reconstructing experimental variograms for five different classes within the evaluated categories using randomly selected realisations. Variogram models were fitted and the variogram parameters (sill, range and nugget) were compared to the error model parameters. From the simulated maps, the classes within category SR/DEC and SR/HEA were used for illustration of the evaluation because (1) these categories display an abundance of several classes, and (2) a wide range of measures of spatial correlation were found within these categories.

Table 1
Areal fraction within each soil/vegetation category on the EU-map occupied by soil/vegetation classes on the NL-map. Empty fields denote no occurence, < 0.01 denotes less than 0.01, ***bold italic*** values indicate purity of EU-category. Soil and vegetation classes are explained in the text

Soil vegetation	EU-category										
	SP DEC	SP HEA	SP GRP	SP CON	SR DEC	SR HEA	SR GRP	SR CON	SC DEC	SC GRP	SC CON
NL-class											
SP DEC	***0.28***	0.05	0.06	0.13	0.21	0.08	0.19	0.13	0.01	0.05	0.10
SP HEA	0.04	***0.62***	0.33	0.07	0.02	0.30	0.30	0.04			
SP GRP	0.03	0.01		0.01	0.05	0.05	0.08	0.03	0.01	0.01	
SP CON	0.51	0.30	0.49	***0.73***	0.30	0.18	0.29	0.62		< 0.01	0.01
SR DEC	0.04	< 0.01	0.01	0.02	***0.14***	0.04	0.01	0.05	0.04	0.03	
SR HEA	< 0.01	< 0.01	0.07	< 0.01	< 0.01	***0.08***		< 0.01			
SR GRP	0.03	< 0.01	0.03	0.01	0.09	0.04	***0.01***	0.03	0.02	0.01	
SR CON	0.03	< 0.01	0.01	0.02	0.07	0.03	0.01	***0.06***	< 0.01	< 0.01	
SC DEC					< 0.01		< 0.01	< 0.01	***0.74***	0.74	0.50
SC HEA							< 0.01				
SC GRP							< 0.01		0.02	***0.02***	0.02
SC CON							< 0.01	< 0.01	0.17	0.14	***0.38***
CN DEC	0.01			< 0.01	0.03	< 0.01	< 0.01	< 0.01			
CN HEA					< 0.01	< 0.01					
CN GRP	< 0.01	< 0.01		< 0.01	0.02	< 0.01	< 0.01	< 0.01			
CN CON	< 0.01	< 0.01		< 0.01	0.01	< 0.01		< 0.01			
CC DEC	< 0.01										
CC GRP				< 0.01	< 0.01			< 0.01			
CC CON											
PN DEC	< 0.01	< 0.01		< 0.01	0.02	0.04	0.02	0.01			
PN HEA	< 0.01			< 0.01	< 0.01	0.11	0.02	< 0.01			
PN GRP	< 0.01			< 0.01	0.01	0.02	0.07	< 0.01			
PN CON	< 0.01	< 0.01		< 0.01	0.01	0.03	< 0.01	0.01			
LN DEC	0.01	< 0.01		< 0.01	0.01	< 0.01		< 0.01			
LN HEA	< 0.01	< 0.01			< 0.01						
LN GRP	< 0.01			< 0.01	< 0.01			< 0.01			
LN CON	0.01	< 0.01		< 0.01	0.01			< 0.01			

3. Results and discussion

3.1. Map impurity

In both the EU-categories and the NL-classes, hypothetically 28 soil–vegetation combinations can occur. When overlaying the SV-EU and SV-NL maps, it occurred that the following soil/vegetation units do not occur in the SV-EU maps: SC/HEA, LN/GRP, PN/GRP and LN/CON. The EU maps seriously underestimate the area of loess soils in the Netherlands. CC/HEA does not

CN DEC	CN HEA	CN GRP	CN CON	CC DEC	CC HEA	CC GRP	CC CON	PN DEC	PN HEA	PN CON	LN DEC	LN HEA
0.17	0.05	0.12	0.14	0.08		0.08	0.12	0.18	0.03	0.16	0.04	0.13
0.02	0.39	0.02	0.02			0.04	0.02	0.02	0.04	0.04	0.01	0.33
0.03			0.02	0.02		0.01	0.01	0.04	0.04	0.02	< 0.01	
0.14	0.43	0.40	0.68	0.01		0.08	0.18	0.13	0.01	0.36	0.05	0.53
0.09	< 0.01		0.03	0.05		0.04	0.01	0.06	0.01	0.05		
	0.01		< 0.01	< 0.01				< 0.01	< 0.01			
0.06			0.01	0.02		0.01	< 0.01	0.01	0.01	< 0.01		
0.02	0.02		0.04	< 0.01			0.01	0.03	< 0.01	0.04		
0.01		0.26	< 0.01	0.13		0.11	0.17	< 0.01				
< 0.01				0.01			0.01					
< 0.01		0.02		0.04		0.06	0.36	< 0.01				
0.19	< 0.01	0.08	0.02	0.13		< 0.01	0.06	0.01			0.23	
< 0.01			< 0.01									
0.13			0.01	0.03		< 0.01	< 0.01	0.03			0.11	
0.01	0.01		***0.01***	0.01			0.01	< 0.01			0.01	
< 0.01				***0.42***	1.00	0.01	0.01				0.01	
< 0.01		0.10		0.02		***0.56***					0.01	
				< 0.01			< ***0.01***					
0.04			< 0.01	0.01		< 0.01	0.01	***0.35***	0.15	0.06	< 0.01	
< 0.01	0.07							0.05	***0.64***	0.07	< 0.01	
0.01				< 0.01			< 0.01	0.06	0.03			
< 0.01	0.01		< 0.01	< 0.01			0.02	0.03	0.05	***0.19***	0.01	
0.05			< 0.01								***0.44***	
< 0.01												
0.01			0.01								0.06	
0.01			< 0.01								0.01	

Table 2
Parameters of 89 fitted exponential variograms of the indicator variable ''Soil/vegetation category s on the EU-map occurs simultaneously with soil/vegetation class t on the NL-map at location x'', with C_0 and C_1 scaled to binominal variance

EU-category	NL-class	Variogram parameters			EU-category	NL-class	Variogram parameters		
Soil vegetation	Soil vegetation	A (m)	C_0	C_1	Soil vegetation	Soil vegetation	A (m)	C_0	C_1
SP DEC	SP DEC	484	0.068	0.134	SC GRP	SC DEC	398	0.102	0.088
SP DEC	SP HEA	2840	0.007	0.032	SC GRP	SC CON	131	0.088	0.034
SP DEC	SP GRP	552	0.012	0.015	CN DEC	SP DEC	432	0.057	0.086
SP DEC	SP CON	1380	0.055	0.195	CN DEC	SP GRP	200	0.016	0.016
SP DEC	SR DEC	100	0.022	0.015	CN DEC	SP CON	419	0.047	0.073
SP DEC	SR GRP	315	0.009	0.022	CN DEC	SR DEC	334	0.036	0.045
SP DEC	SR CON	3880	0.009	0.020	CN DEC	SR GRP	300	0.024	0.034
SP HEA	SP DEC	80	0.023	0.023	CN DEC	CN DEC	137	0.057	0.096
SP HEA	SP HEA	630	0.034	0.201	CN DEC	CN GRP	200	0.046	0.064
SP HEA	SP CON	603	0.065	0.145	CN DEC	PN DEC	140	0.011	0.027
SP CON	SP DEC	421	0.044	0.066	CN DEC	LN DEC	597	0.003	0.048
SP CON	SP HEA	1590	0.007	0.060	CN CON	SP DEC	151	0.045	0.075
SP CON	SP CON	715	0.049	0.148	CN CON	SP CON	369	0.000	0.217
SP CON	SR CON	83	0.012	0.012	CN CON	SR DEC	209	0.011	0.020
SR DEC	SP DEC	150	0.077	0.087	CN CON	SR CON	766	0.024	0.019
SR DEC	SP GRP	200	0.025	0.025	CC DEC	SP DEC	100	0.047	0.023
SR DEC	SP CON	733	0.085	0.123	CC DEC	SR DEC	100	0.023	0.023
SR DEC	SR DEC	100	0.025	0.098	CC DEC	SC DEC	200	0.043	0.070
SR DEC	SR GRP	207	0.011	0.071	CC DEC	SC CON	200	0.005	0.038
SR DEC	SR CON	261	0.010	0.056	CC DEC	CN DEC	150	0.039	0.077
SR DEC	CN DEC	100	0.005	0.020	CC DEC	CN GRP	150	0.000	0.029

SR HEA	SP DEC	92	0.050	0.025	CC DEC	CC DEC	1210	0.009	0.234
SR HEA	SP HEA	337	0.046	0.165	PN DEC	SP DEC	422	0.066	0.081
SR HEA	SP GRP	1320	0.019	0.031	PN DEC	SP GRP	1000	0.021	0.021
SR HEA	SP CON	379	0.053	0.096	PN DEC	SP CON	2210	0.036	0.080
SR HEA	SR DEC	307	0.014	0.021	PN DEC	SR DEC	824	0.009	0.046
SR HEA	SR HEA	543	0.011	0.061	PN DEC	SR CON	2140	0.000	0.029
SR HEA	SR GRP	1230	0.006	0.032	PN DEC	CN GRP	100	0.000	0.028
SR HEA	SR CON	439	0.009	0.016	PN DEC	PN DEC	385	0.043	0.183
SR HEA	PN DEC	55300	0.015	0.023	PN DEC	PN HEA	800	0.010	0.040
SR HEA	PN HEA	572	0.024	0.072	PN DEC	PN GRP	200	0.014	0.041
SR HEA	PN CON	527	0.012	0.014	PN DEC	PN CON	400	0.007	0.021
SR CON	SP DEC	157	0.077	0.037	PN HEA	SP DEC	1880	0.012	0.014
SR CON	SP HEA	438	0.006	0.031	PN HEA	SP HEA	294	0.012	0.024
SR CON	SP GRP	500	0.016	0.016	PN HEA	SP GRP	1290	0.017	0.021
SR CON	SP CON	353	0.006	0.230	PN HEA	PN DEC	420	0.062	0.064
SR CON	SR DEC	186	0.024	0.024	PN HEA	PN HEA	371	0.046	0.184
SR CON	SR GRP	200	0.000	0.029	PN HEA	PN GRP	443	0.008	0.020
SR CON	SR CON	483	0.020	0.038	PN HEA	PN CON	115	0.014	0.031
SC DEC	SR DEC	694	0.021	0.014	LN DEC	SP DEC	100	0.005	0.030
SC DEC	SC DEC	340	0.070	0.122	LN DEC	SP CON	200	0.000	0.044
SC DEC	SC CON	200	0.074	0.065	LN DEC	CN DEC	350	0.046	0.132
SC GRP	SP DEC	206	0.008	0.037	LN DEC	CN GRP	199	0.020	0.076
SC GRP	SR DEC	1100	0.003	0.025	LN DEC	LN DEC	163	0.080	0.167
					LN DEC	LN GRP	193	0.031	0.028

occur only on the SV-NL map. Table 1 summarizes the areal fraction p within each EU-category occupied by an NL-class. The purity of the EU-maps (the fraction of EU-category where NL-class = EU-category; bold italic values in Table 1) varies greatly between 0% and 74%. Some figures point to apparent misclassifications in the EU-map, e.g., all heather vegetations in calcareous clay soils in SV-EU are actually deciduous forests in SV-NL. The latter makes much more sense, since heather vegetations in the Netherlands originate from extensive grazing on poor sandy soils and do not occur on fertile clay soils. Many noncalcareous clay soils in SV-EU are actually poor sandy soils in SV-NL.

The overall-purity of the EU-maps equals 0.31, so 69% of the area is misclassified when compared to highly detailed NL-data. Based upon map purity estimates from the NL soil map and landcover map and assuming mutual independence, the purity of SV-NL would be roughly the product of the soil and vegetation map purity: $0.7 \times 0.9 = 0.63$. This assumption is probably invalid, because in the Netherlands (semi)natural vegetations are usually associated with poorer (sandy) soils. The actual purity would probably be between 0.63 and 0.70. Forgoing this invalidity and assuming independence between SV-EU and SV-NL too, the actual purity of SV-EU might be $0.31/0.63 = 0.45$, which is still considerably lower than SV-NL. SV-EU and SV-NL can safely be assumed independent, because the production of both maps took place independently and the base material for SV-EU existed before that of SV-NL (SV-EU is no generalization of SV-NL). Therefore, it can safely be assumed that a misclassification at any point in SV-NL has no relation to a possible misclassification at the same point in SV-EU. Because a precise quantification of the map impurity of SV-NL was beyond the scope of this study, we treated the SV-NL maps as if they were perfect, and did not correct differences between SV-EU and SV-NL for errors in SV-NL.

3.2. Spatial correlation

Exponential variograms were fitted to indicator variables relating to the larger areas on both the SV-EU and SV-NL maps. To avoid the work load of modelling variograms for SV-EU/SV-NL combinations that hardly ever occur, indicator variograms were only made for EU-categories occupying more than 1500 ha combined with NL-classes larger than 1500 ha also or occupying more than 2.5% of the area within this EU-category. Accordingly, 89 variograms were fitted, representing 97.8% of the area within SV-EU (Table 2).

The values of the range-related parameter A show that autocorrelation exists up to distances of between 240 m and several kilometers, since in exponential models, the practical range ≈ 3 A (Journel and Huijbregts, 1978). The uncertainty about impurities in the SV-EU map does not increase beyond distances of a few kilometers, so if cell sizes are large enough, uncertainty inside these cells levels out. Aggregating of the output of the SMART2-model to cell sizes of a

few square kilometers therefore results in between-block variances that are mainly due to ''true'' soil variability. The coefficients used to correct fitted $(C_0 + C_1)$ to binominal variances vary between 0.8 and 14.7. The higher scaling coefficients (larger than 3) are nearly always accompanied with sample variograms that look more like hole-effect models than exponential models. In these cases (23 out of 89 fits), poor exponential fits can be expected, which did result in underestimated sills.

3.3. Map realisations

In Figs. 1 and 2, the observed mean areal fractions, expressed as percentages, for each class within category SR/DEC are compared. According to the error model, 22 different classes are possible within this category. For the 22 classes, the abundance ranges from 0.03% to 29.7%. In Figs. 1 and 2, a cross without a corresponding circle indicates that the simulation did not produce any cell of that class. Simulations 4 or 7 do not reproduce either two or four of the 22 NL-classes within the EU-category SR/DEC. This is due to the low abundance of these classes in the error model. The areal fractions of the other NL classes are reproduced nicely by the simulations.

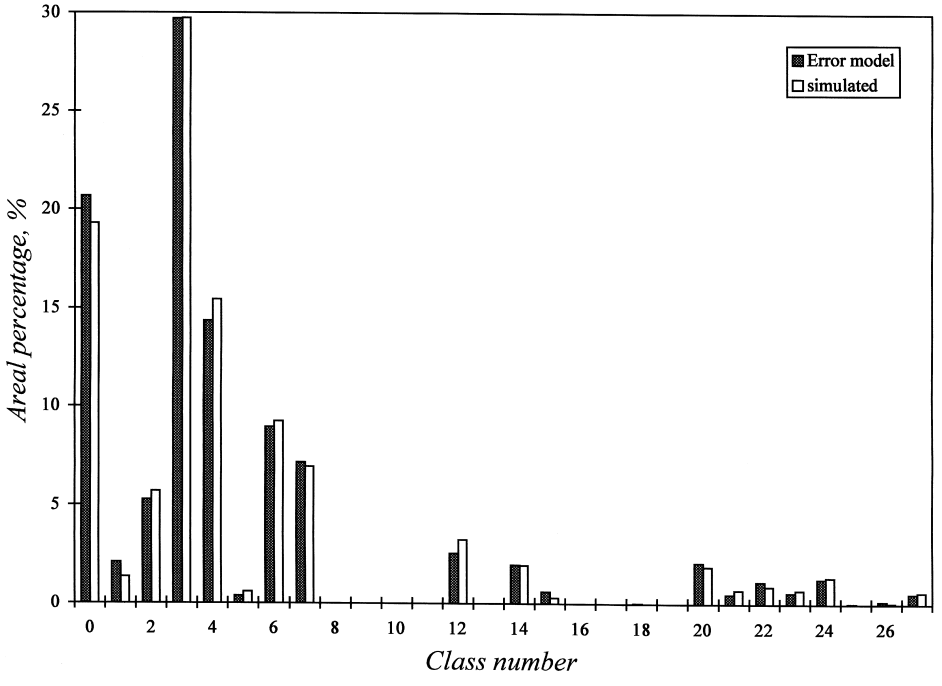

Fig. 1. The simulated vs. the measured areal fraction for the 22 soil/land use classes within category SR/DEC for simulation 4.

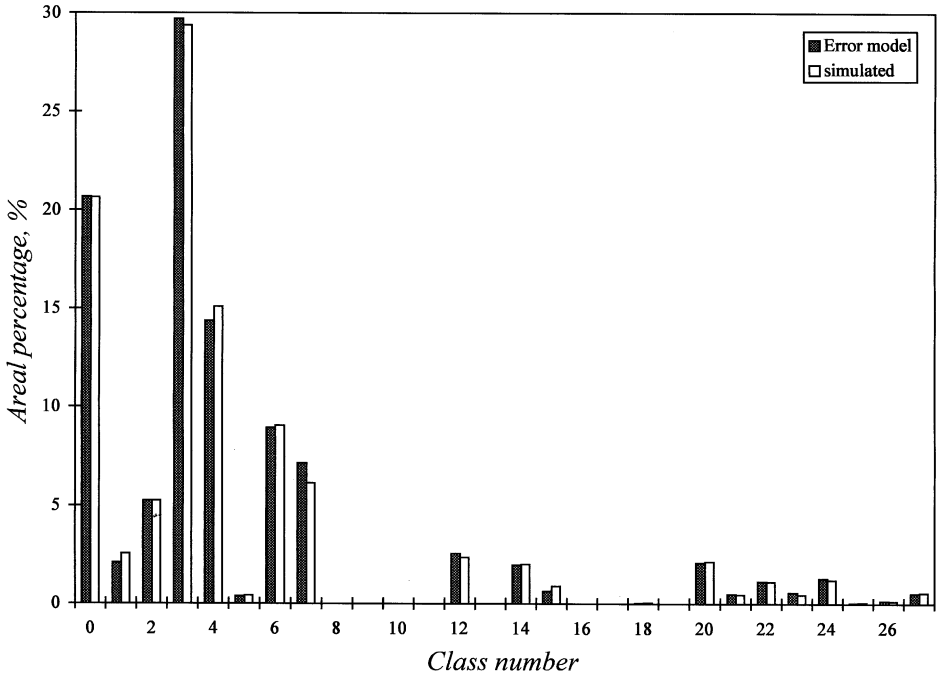

Fig. 2. The simulated vs. the measured areal fraction for the 22 soil/land use classes within category SR/DEC for simulation 7.

The spatial variability of the simulations was evaluated by comparing experimental variograms from simulated data to the fitted variogram for the error model. A number of randomly chosen simulations was plotted and the error model variogram was merged into the plot. Due to the resolution of simulation data (1 km $\times$ 1 km), a practical range larger than 1 km was necessary for this evaluation.

Fig. 3 gives the experimental variograms for five simulations for the class SP/CON within the category SR/DEC. The range and the mean are determined for the error model to 733 m and 29.7%. As described in the previous section, the practical range for the error model can be determined to 2200 m. The five selected simulations show close agreement with the error model as far as the sill variance is concerned. The resolution of the simulated maps does not allow a comparison between simulated and error model semivariances at lags much smaller than the practical range. Likewise, Fig. 4 displays the class SP/GRP within the category SR/HEA. The range and the mean are determined by the error model to 1320 m and 5.3%, respectively. The experimental variograms based on simulated data indicate too small an abundance of the class SP/GRP to be able to reproduce the error model variogram. A small areal abundance for a class results in erratic estimates of experimental semivariances.

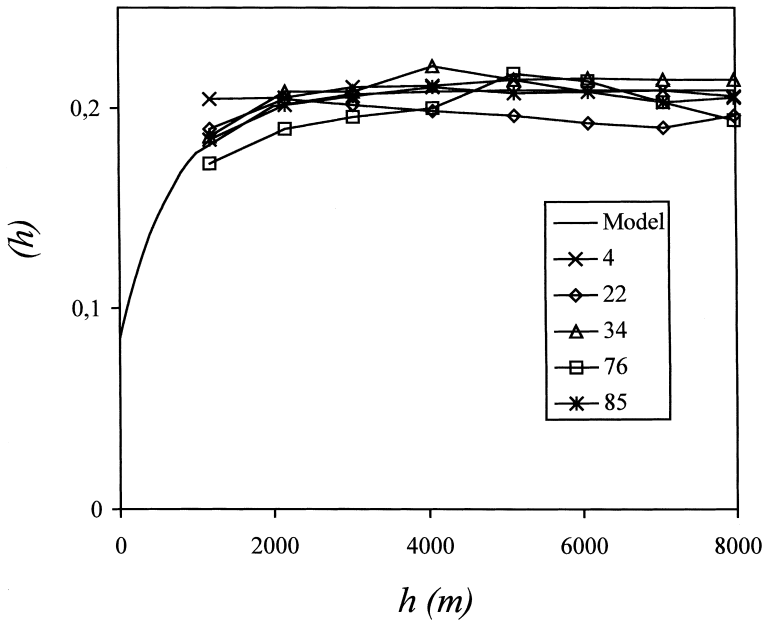

Fig. 3. Experimental variograms for five realisations of class SP/CON within category SR/DEC (marked lines) and the error model variogram.

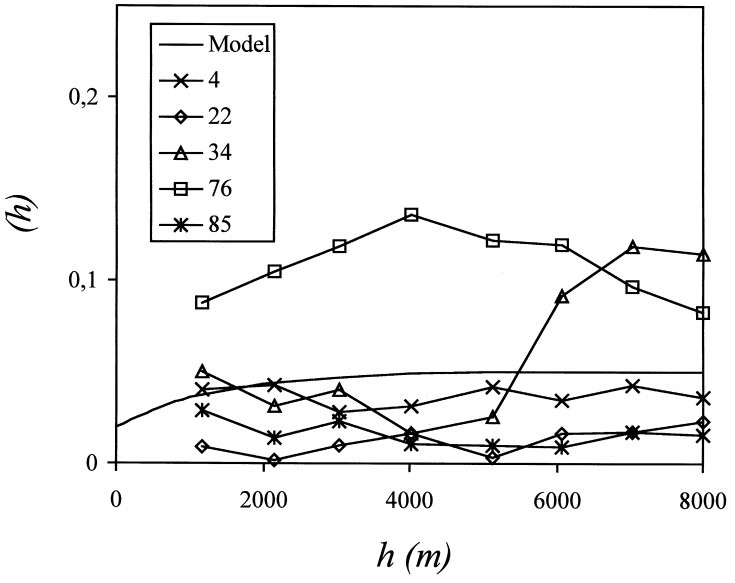

Fig. 4. Experimental variograms for five realisations of class SP/GRP within category SR/HEA (marked lines) and the error model variogram.

3.4. Uncertainty analysis

To quantify the relative contributions of uncertainty in categorical data, in continuous soil parameters and in continuous vegetation parameters, an analysis of variance was conducted on block-aggregated model outputs. A complete description of the procedure, including the simulation of continuous parameters, is given by Kros et al. (1999). In summary, for each of the 25 realisations of the soil/vegetation map, five realisations of the soil parameters were combined with five realisations of the vegetation parameters, so the Monte Carlo analysis comprised in total $25 \times 5 \times 5 = 625$ input maps for each of the 18 continuous soil and vegetation-related input parameters to SMART2. One map contains 7435 1 km $\times$ 1 km grid input locations in The Netherlands. As exogenous inputs, the European coordination scenario was used, which is an official deposition scenario defined in the Dutch Environmental Outlook (RIVM, 1997) with data sets for the years 1995, 2000, 2010 and 2020. The following model output parameters were considered: (i) Al^{3+} concentration below the root zone, (ii) probability that the Al^{3+} concentration exceeds a threshold of 0.2 $mol_c\ m^{-3}$ (a forest vitality threshold value), (iii) probability that the Al^{3+} concentration exceeds a threshold of 0.02 $mol_c\ m^{-3}$ (maximum allowable drinking water

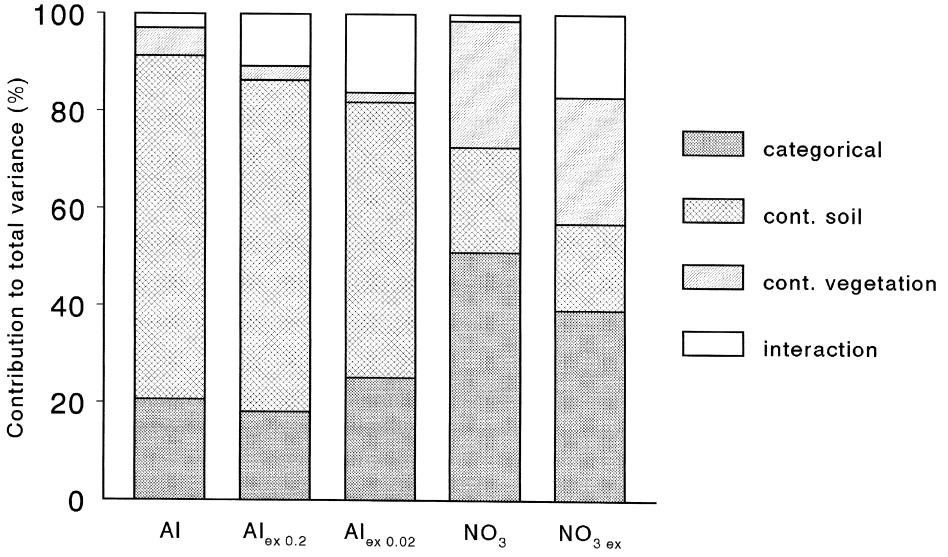

Fig. 5. Relative variance contributions in five model output parameters by four sources of uncertainty in model inputs. Model outputs relate to Al^{3+} and NO_3^- concentrations and concentration thresholds in phreatic groundwater for 5 km $\times$ 5 km blocks. Uncertainties of model inputs concern categorical soil/vegetation data (dark gray), continuous soil data (fine crosshatch), continuous vegetation data (light gray) and their interactions (blank).

concentration), (iv) NO_3^- concentration below the root zone, (v) probability that the NO_3^- concentration exceeds a threshold of 0.81 $mol_c\ m^{-3}$ (i.e., 50 mg/l, the maximum allowable concentration for drinking water within the EU). Outputs were aggregated to 5 km × 5 km blocks before analysis.

Results of the analysis for the year 1995 are given in Fig. 5 (taken with permission from Kros et al., 1999). Fig. 5 shows clearly that uncertainty contributions from continuous soil data dominate the total uncertainty with respect to the Al^{3+}-related outputs, since variance contributions of these data vary between 55% and 70% of total variance. The effect of the uncertainty in soil vegetation maps comprises about 20% of the total variance, where the most serious contribution occurs in areas with low Al^{3+} concentrations, probably due to misclassification of the calcareous soils in the SV-EU map (Kros et al., 1999). For the nitrate-related model outputs, however, the effect of the uncertainty in the SV-EU map is dominant (40%–50%), and is comparable to the sum of the influences of uncertainty in continuous soil and vegetation data and their interactions.

4. Conclusions

Soil and vegetation maps that are used for EU-wide soil acidification risk assessment show a high level of impurity. Up to 69% of the area of The Netherlands is misclassified when compared to highly detailed soil and vegetation maps.

To quantify the effect of the uncertainty in these categorical data on the uncertainty of the soil acidification model SMART2, we successfully applied a method essentially comprising the following steps: (i) the construction of an error model describing the degree of misclassification and its spatial autocorrelation, and (ii) the drawing of map realisations by sequential multiple indicator simulation. Subsequent steps have been described elsewhere and involve (iii) drawings from the error distributions of continuous model input parameters and (iv) an analysis of variance to identify the effect of different sources of uncertainty.

The current study shows that errors in categorical data due to misclassification and scale do have a pronounced influence on the uncertainty of the results of SMART2. This influence varies with the output parameter being considered. The uncertainty due to categorical data could be reduced by (i) validation of the classification of the categories in the SV-EU-maps, since many apparent misclassifications exist, and (ii) substituting the SV-EU maps with data from more detailed sources, like the 1:250,000 soil maps which are available for considerable parts of Europe.

Acknowledgements

This work was sponsored by the EU, Project ENV4-CT95-0070, UNCERS-DSS. We thank F. de Vries and J.C. Voogd of SC-DLO for the GIS-operations and Erik Larsson of Chalmers University for the assistance with the simulations.

References

Deutsch, C.V., Journel, A.G., 1992. GSLIB: Geostatistical Software Library and User's Guide. Oxford Univ. Press, New York, 340 pp.

De Vries, W., de Kros, J., Salm, C., van der Groenenberg, J.E., Reinds, G.J., 1998. The use of upscaling procedures in the application of soil acidification models at different spatial scales. Nutrient cycling in Agroecosystems 50, 223–236.

EC, 1985. Soil map of the European Communities, 1:1,000,000. Commission of the European Communities, Office for Official Publications and Directorate General VI — Agriculture of the European Communities, Luxembourg.

EC, 1993. Corine land cover. Technical guide. Office for Official Publications and Directorate General Environment, Nuclear Safety and Civil Protection of the European Communities, Luxembourg.

Goméz-Hernández, J.J., Journel, A.G., 1992. Joint sequential simulation of multigaussian fields. In: Soares, A. (Ed.), Proceedings of the Fourth Geostatistics Congress Troia (Portugal), Quantitative Geology and Geostatistics 5 Kluwer Academic Publishers, pp. 85–94.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, 483 pp.

Hoosbeek, M.R., Bryant, R., 1992. Towards the quantitative modelling of pedogenesis — a review. Geoderma 55, 183–210.

Jansen, M.J.W., 1998. Prediction error through modelling concepts and uncertainty from basic data. Nutrient Cycling in Agroecosystems 50, 247–253.

Journel, A.G., Huijbregts, Ch.J., 1978. Mining Geostatistics. Academic Press, London.

Kleeschulte, S., 1997. In: Assessment of ''cumulative'' uncertainty in Spatial Decision Support Systems. Proceedings of 3rd EC-GIS Workshop, Leuven, Belgium, 25–27 June 1997. .

Kros, J., Pebesma, E.J., Reinds, G.J., Finke, P.A., 1999. Uncertainty in modelling soil acidification at the European scale, a case study. Journal of Environmental Quality 28, 366–377.

Noordman, E., Thunnissen, H.A.M., Kramer, H., 1997. Construction and Accuracy of the LGN2 Land Cover Database. DLO Winand Staring Centre, Rapport 515, Wageningen, (in Dutch).

Pebesma, E.J., Wesseling, C.G., 1997. Gstat, a program for geostatistical modelling, prediction and simulation. Computers and Geosciences 24, 17–31.

RIVM, 1997. National Environmental Outlook 1997–2020. RIVM, Bilthoven, (in Dutch).

Steur, G.G.L., Heijink, W., 1991. Soil map of The Netherlands at scale 1:50,000. General Concepts and Structure. 4th edn. DLO-Staring Centrum, Wageningen, (in Dutch).

Thunnissen, H.A.M., Olthof, R., Gertz, P., Velts, L., 1992. Land cover database of The Netherlands based on Thematic Mapper images. DLO Winand Staring Centre, Rapport 168, Wageningen, (in Dutch).

Wösten, J.H.M., Finke, P.A., Jansen, M.J.W., 1995. Comparison of class-and continuous pedotransfer functions to generate soil hydraulic characteristics. Geoderma 66, 227–237.